

Characterizing the nature of interactions for cooperative creation in online social networks

Remy Cazabet¹  · Hideaki Takeda¹ · Masahiro Hamasaki²

Received: 12 March 2015 / Revised: 6 July 2015 / Accepted: 9 July 2015 / Published online: 21 July 2015
© Springer-Verlag Wien 2015

Abstract Many aspects of online social networks (OSN) have been studied in recent years. In this article, we focus on the question of interactions in large OSN. We propose methods to study these interactions, and apply them on a platform called Nico Nico Douga (NND), with the aim of understanding cooperative behaviors, taking the form of collective creation of music videos in NND. Our first contribution is a method that, starting from the network of interactions between users, evaluates three aspects: the impact of the social structure on these interactions, their concentration, and their reciprocity. We characterize the nature of interactions in NND, and compare it with four different datasets. We find that interactions in NND are more similar to a diffusion process, such as retweets in Tweeter, than to interpersonal communications, or even to cooperation in science. Our second contribution is a typology of roles for productions in a cooperative process. These roles are attributed based on the neighborhood of the nodes in the network of references between productions. We define direct roles, relative roles, and indirect roles. We subsequently study the frequency of these roles in NND. We show a correlation between the category of the contribution of a video (song, animation, etc.) and its probability of having a certain role. We also find a positive correlation between the most active users and the

production of videos playing an important role in the cooperation process.

Keywords Social network analysis · Mass cooperation · Nico nico douga · Artistic cooperation · Interaction network

1 Introduction

Online social networks (OSN) have attracted a lot of attention from scientists of many different fields. By the large quantity of data about human behaviors that they make available, they allow to study interactions between individuals in a way that was not possible before. Among the tremendous amount of work published, we can cite works on information diffusion such as (Bakshy et al. 2011; Yang and Counts 2010), structural properties (Amaral et al. 2000; Clauset et al. 2009), community structures (Leskovec et al. 2009), influence (Cha et al. 2010), and so on and so forth. A wide variety of networks have been studied, including Facebook, Twitter, Wikipedia, and many others. In this paper, we focus on an OSN called Nico Nico Douga (NND, also known as Niconico outside Japan), one of the most popular in Japan, which has already been studied in a few articles (Hamasaki et al. 2008; Nakamura et al. 2008; Cazabet et al. 2012; Hamasaki et al. 2009). We will briefly present this platform in the next section.

This article focuses on the interactions between the constitutive elements of a social network: actors and published entities. We can find both of these elements in most social networks, sometimes with a focus on actors (such as in Facebook), and sometimes on the content (Wikipedia for instance). Nico Nico Douga is an interesting platform on

✉ Remy Cazabet
remy.cazabet@gmail.com

¹ National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-8430, Japan

² National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan

this aspect, because we have a lot of raw data both on the creators of videos (actors) and on the videos (their published creations). However, little is known about their interactions. The nature of these interactions is especially interesting among NND users involved in the cooperative creation of music videos. These users work together, in a decentralized manner, to create complex productions that often reach a large audience. Such large-scale artistic cooperation is little studied despite being an interesting phenomenon.

After the introduction of the NND platform and our dataset, we will present a method to study the characteristics of interactions between actors in a social network that we apply to NND and several other better-known social interaction datasets. By comparing the results obtained, we can classify the interactions between authors in NND as similar to those happening in a Twitter retweet dataset. The second part of the paper focuses on the interactions between the productions published on the network, which are videos. We present several roles, defined by the network of references between productions. We study on NND how different categories of videos have different roles.

2 Nico Nico Douga network

Nico Nico Douga¹ (NND) is a video sharing social network, with functionalities comparable to those of YouTube or DailyMotion. NND originated from Japan, and is extremely popular in this country, with over 20 Million registered users as of 2014, and ranking in the top 15 of the most visited websites. Compared to the previously cited platforms, NND offers some additional possibilities that we will use in this paper:

- Users can associate free keywords with videos. Keywords can be associated or removed by any registered user to any video. Some keywords, set by the author of the video, cannot be changed. These keywords are used by the platform to group the videos into categories (music, animation, sport, etc.). We use all these keywords to discover the categories of videos.
- For each video, it is possible to add references to other videos. It is a common practice in NND to list, in the description written by the author for a video, the ID number of all videos that have been used for its creation. The author of the video is the only person able to write in this comment page, and can edit it at any time to complete it, if needed. We use these references to create a network of interactions between videos.

Additionally to these functionalities, NND provides some social aspects. First, each user has a webpage, where one can find all its uploaded videos, and a list of its favorite videos. Furthermore, a wiki called NicoNicoPedia is directly integrated into the platform, and one can access and edit explanations associated with famous creators, keywords, or videos. There is a strong community of active users around NND, which has strongly contributed to its success. One specificity of NND that we do not use in this article is that users can add comments directly on the video, that appears overlaid on the video, at the time and position chosen by the author of the comment. Finally, users can find videos on the website by several means. First, NND highlights some currently popular videos, it is also possible to search for keywords, authors, text in the comments, etc. Results can be sorted by number of views, date of publication of the video, number of comments, date of most recent comment, or number of times this video has been chosen as a favorite by other users.

Figure 1 gives an example of the page associated with a video.

2.1 NND dataset

The dataset that we use has been described in previous papers (Hamasaki et al. 2013; Cazabet and Takeda 2014). It has been constituted by crawling a set of metadata associated with all the videos published on the network between January 2007 and December 2012. It is composed of a set of 2.6 Million videos with at least one keyword associated with them. For each video, metadata consist of their author, associated keywords, associated description (author comment), and date of publication.

As we are interested in relations between videos and authors, we focus on the very important phenomenon of collaborative music video creation on NND.

2.2 VOCALOID, hatsune miku, and music videos

VOCALOID is a singing voice synthesizer, a special voice synthesizer that not only able to pronounce words but also to sing them according to a defined tune. Since its introduction in 2004, this software has encountered a huge popularity, particularly in Japan. NND played an important role in this popularity. Some users first created songs and published them on NND as music videos, usually with very simple visuals such as a static drawing. Other users liked these songs, and started to produce derived music videos, modifying the visuals, the voice, the music, in thousands of different ways. Inspired by the character represented on the packaging of the software, they started to assimilate all songs composed with the synthesizer with a fictional singer, called Hatsune Miku. Although other voices have

¹ <http://nicovideo.jp>

Fig. 1 Page of a video in NND

been created for VOCALOID, corresponding to different characters, Hatsune Miku remains the most famous one. While the authors of the songs and derived videos were initially amateurs, they became so famous that some of them started to release commercial versions of their productions. For instance, the group of producers known as Supercell have published albums sold in hundreds of thousands of copies.

It is this cooperative creation of music videos that we will study in this article. To give an idea of the scale of the phenomenon, more than 200,000 videos have been tagged with the keyword VOCALOID in our dataset, and more than 130,000 with the keyword Hatsune Miku.

2.3 Categories of videos

The productions in NND can be of several types, and we wanted to study the similarities and differences between

these types. By our knowledge of the network and observation of common tags, we derived a classifier that, according to the keywords associated with a video, attributes a category to it, using direct matching and regular expressions. The possible categories are the following:

- **OriginalMusic**: an original musical composition
- **Singing**: a person is singing (example: replace the original voice of a famous song)
- **VocaloidVoice**: a person is using VOCALOID to create a voice (example: replace the original voice of a famous song)
- **MusicalPerformance**: a person uses mMusical instruments to create this video (example: add an instrument to a famous Music)
- **Picture**: a user creates one or several static pictures (example: illustrate a famous song with original drawings)
- **Dance**: a user films himself dancing on a famous song)

- **3DCG**: a user uses a 3D Computer Graphic software to create this video (example : animate dancers dancing on a famous music)
- **Animation**: a user creates an animated picture (example : illustrate the lyrics of a famous song)
- **Mashups**: a mashup is a music video created by combining several original sources, such as different original musics or, more often, different versions of a same music.
- **MAD**: MAD videos are an original type of video originally invented in Japan, involving a collage of videos and sounds from multiple sources. Compared to **Mashups**, **MAD** videos are more diverse, as they can be composed of people speaking, unrelated sounds or pictures, and do not usually compose a single, coherent music or song.
- **Movie**: the content is a movie illustrating a song, without other precision. It can be a different editing of an existing video, or an existing video with modifications such as the addition of special effects, addition of a contour frame, or hue alteration for instance.
- **Voice**: a user modifies the vocals of another video. It can be the application of sound filters for instance, however the most common case is the creation of *karaoke videos* where the voice is removed (but the music remains).

2.4 References

References between these videos have been collected: for each video, the associated description by the author has been crawled, and references to other videos have been identified. As we explained previously, it is a common practice to cite other videos this way, in particular among music video creators. A total of 7.9 million references have been identified this way. One problem is that these references can have different usages. For instance, they can be used to refer videos in a same series by the same author, as a long video cut in several parts, or just to link the other creations by the same authors. As we focus on interactions, we filtered out all references made from a video to a later one (authors can add references after the publication of the video), and all references between two videos created by the same author. Although this might suppress some interesting references, it allows us to focus on actual cooperation between individuals. The resulting graph is composed of approximately 800,000 nodes and 1.1 million edges, following a typical long-tailed distribution for the degrees, as illustrated in Fig. 2. Note that there are some irregularities in the distribution for the out-degrees around $d = 20$ and $d = 40$, probably due to a platform limitation.

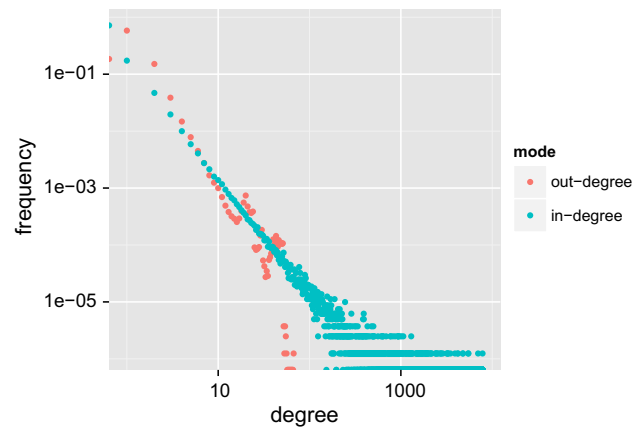


Fig. 2 Degree distribution of the Nico Nico Dataset

3 Structure of the interactions

In this section, we study what is the influence of the social structure existing among actors on the way these actors interact with each other.

In NND, there is no explicit declaration of social relations between individuals, unlike OSNs such as Facebook or Twitter, which have “Friends”, “Followers”, or “Followees”. We do not know if authors cooperate with users with who they have personal bonds, such as friendship bonds, or if most references are made to a minority of celebrities, for instance.

We propose a method to evaluate the strength and the nature of social relations among a group of users by studying their interactions. We define the network of interactions as $G = (E, V)$ an oriented multigraph (a graph that can contain multiple edges between the same nodes, called parallel edges). Nodes correspond to actors, and edges to interactions between these actors. An interaction is not necessarily reciprocal: if a sends a message to b , we consider it as an interaction between a and b . The order of the interactions is not considered. If a sends three messages to b and b sends two messages to a , this will be represented in G by three directed edges from a to b and two from b to a . In the case of NND, nodes are videos’ creators, and an edge (a, b) corresponds to the publication by a user a of a video v containing a reference to a video v' published by user b .

One possible approach to such a problem could be to identify which pairs of users are involved in a social relation and which aren’t. Although this approach would have the advantage of providing us with a network that can subsequently be studied by the mean of usual, topological approaches, such as degree distributions or clustering measures, the main problem of such a network is that it cannot be constructed in a reliable manner. The reason is that, in order to construct a topological network from

sequences of interactions, it is necessary to identify which communications can be considered the consequence of social relations between users, and which ones cannot, because they could also have occurred if interactions were randomly distributed among actors, or due to noise in the dataset (for instance, messages sent to the wrong recipients, occasional messages sent in behalf of someone else, hacked account, etc.). To identify these significant social relations to be represented by a bond, we would have to set a threshold above which interactions are considered important enough to infer the existence of a social bond underlying them. This threshold can be either static—the same for any pair of node—or dependent on the properties of the nodes—the most active users are more likely to have repeated interactions among them. Both methods have their own drawbacks. With a fixed parameter, there will be a bias towards active users. If the parameter is chosen with a low value—of two, for instance—any repeated interaction is considered significant, but the most active users will be able to reach this threshold too easily. If the value chosen is high on the contrary, the bulk of normal users will not have enough observed interactions to reach it. If the threshold is chosen dynamically according to the activity of users, the problem is that, in very large networks, the probability for any pair of users to interact randomly is so low, that any interaction observed will be considered significant; and if one do not wish to consider each and every communication as significant, one will have to set up parameters or thresholds, with the same drawbacks as previously mentioned.

The solution we propose is to work on topological networks, but never to look at the details of these networks, such as who is friend with who, but instead only to look at global properties, and to compare these global properties to the ones of null models, to determine how much of it is due to the nature of this particular network, and how much of it can be explained by chance. We propose to use this technique to compute three aspects of the effects of the social structure on interactions, namely the social structure impact (SSI), the reciprocity impact (RI), and the concentration impact (CI).

3.1 Computation of the metrics

Each metric is computed by going through three steps:

1. Generation of a null model
2. Computation of the frequency distribution of the studied network property
3. Computation of the metric by comparing the distributions corresponding to the observed network and the null model.

Step 3 is the same for all metrics, and is therefore described a single time for all, in the next section. Steps 1 and 2, on

the contrary, depend on the metric to compute, and are described in the corresponding sections.

3.2 Comparing frequency distributions

The question we want to answer by comparing an observed frequency distribution and the same distribution in a rewired version of the graph is: “how much of the distribution can be explained by randomness”. The answer to this question is provided as a value between 0 and 1, corresponding to the fraction of the observation that can be explained by the null model. As an example, Fig. 3 represents the distribution of repeated interactions on an observed network and its corresponding null model. The area corresponding to unexplained observations is filled in gray.

For an observed discrete distribution $D_o(x) = y$ and a distribution obtained from a null model $D_n(x) = z$, the unexplained difference ratio UDR for a value $x \in \mathbb{N}$ is defined as:

$$\text{UDR}(x) = \begin{cases} 0 & \text{if } x = 0 \\ \max(0, \text{UDR}(x-1) + ((D_o(x) - D_n(x)) * x)) & \text{if } x \geq 1 \end{cases}$$

The total difference is computed as $\text{UDR}(m)$, where m is the maximal value such as $D_o(m) > 0$ and $D_n(m) > 0$.

3.3 Social structure impact (SSI)

The first aspect that we investigate is how much the social structure of users affects their interactions. We use social structure as a broad term that includes social relations, as

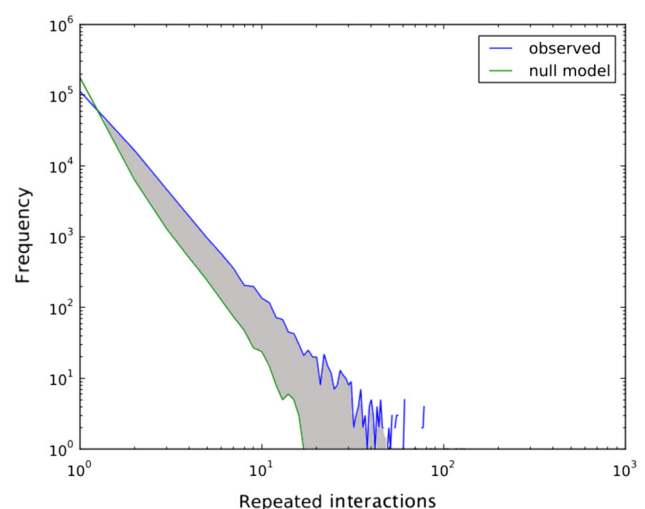


Fig. 3 Distributions of multiple interactions for an observed dataset and its corresponding null model. The gray area corresponds to the unexplained observations

usually understood in the field of social networks, but also as any kind of influence that could lead to a bias in the interactions, such as the impact of language, shared center of interest, or geographical location.

3.3.1 Generation of the null model

The null model corresponding to this multigraph is obtained by generating a random network with the same in-degree and out-degree distribution as the observed interaction multigraph. These degrees correspond to the behavior of users in terms of interactions. They have a non-negligible effect on the probability of having multiple interactions between users. In particular, skewed distributions tend to increase the number of multiple interactions between the same users.

3.3.2 Computation of the distributions of multiple interactions

For both the observed network and the null model, we compute the distribution of multiple interactions. The number of multiple interactions for an ordered pair of nodes (a, b) in a multigraph $g = (V, E)$, noted $\text{mul } I(g, a, b)$, is equal to: $\text{mul } I(g, a, b) = |\{e \in E | e = (a, b)\}|$. The distribution for a multigraph $g = (V, E)$ at a value i is then defined as

$$D(g, i) = |\{(a, b) | \text{mul } I(g, a, b) = i, a \in V, b \in V\}|$$

An illustration of this distribution for an observed dataset and its corresponding null model is represented in Fig. 3

3.3.3 Computation of the social structure impact

The social structure impact is computed as the difference between the observed frequency distribution of repeated interactions and the one in the null model, as explained in section 3.2.

3.4 Reciprocity impact (RI)

In the previous section, we proposed a method to compute the impact of the social structure in general on the interactions between actors. Another characteristic that we can study is the amount of reciprocity existing in the interactions among these actors.

For instance, we expect high values of reciprocity in conversational communications, while a low value is expected for communications corresponding to diffusion of information.

We take the same approach as previously: we compare the distribution of reciprocal messages between the observed multigraph and a randomized version of it.

3.4.1 Distribution of reciprocal messages

The number of reciprocal messages for an ordered pair of nodes (a, b) , in a multigraph $g = (V, E)$, noted as $\text{reci}(g, a, b)$, is defined as the smallest value between the number of parallel edges (a, b) and the number of parallel edges (b, a) , more formally: $\text{reci}(g, a, b) = \min(|\{(a, b) | (a, b) \in E\}|, |\{(b, a) | (b, a) \in E\}|)$. It is a simple measure of the reciprocity in the interactions between these nodes, that do not consider their order, but simply their quantity. The distribution for a network $g = (V, E)$ at a value i is then defined as

$$D\text{reci}(g, i) = |\{a, b, \text{reci}(g, a, b) = i\}|$$

3.4.2 Generation of the null model

The null model that we use in this case is the same as the one used to compute the Social Structure Impact.

3.4.3 Computation of the reciprocity impact

We compute the difference between the distribution of reciprocal messages in the null model and in the observed model as described in section 3.2.

3.5 Concentration impact (CI)

The aim of this indicator is to measure how important is the concentration of interactions. Depending on the type of interaction studied, it is possible that the amount of messages received by actors is evenly distributed, or, on the contrary, a few actors might concentrate most of them. We know that social networks in general tend to follow power law distributions, and, therefore, show great concentration. However, different usages have different concentration profiles. The indicator varies between 0 and 1, 0 corresponding to a minimum concentration—a random distribution of communications—while 1 corresponds to a maximum concentration—none of the concentration observed can be explained by the random model.

Note that the null model used takes into account the concentration of interactions due to friendships, as we explained below.

3.5.1 Distribution of received messages

We compute the distribution of the number of interactions received per user. For each user, this value corresponds to its in-degree in the interaction multigraph. It represents how popular is this user, in terms of the number of interactions he receives. Note that this value for an actor a depends both on how many different users interact with a ,

and of the amount of interactions between these users and a . Therefore the received messages for a user a in a multigraph $g = (V, E)$, $\text{rec}(g, a)$, is equal to:

$$\text{rec}(g, a) = |\{(b, a) | (b, a) \in E, b \in V, \}|$$

and the distribution of received messages is defined as:

$$D_{\text{rec}}(g, i) = |\{x | \text{rec}(g, x) = i, x \in V\}|$$

3.5.2 Generation of the null model

The null model that we use in this case is different from the previous one. Our aim is to estimate how the distribution of in-degrees would be like if the destination of interactions were chosen randomly. We create the random multigraph corresponding to the null model by rewiring the original multigraph. However, we want to keep one property of the original: the distribution of the number of interactions between pairs of vertices. Because of preferences between users such as friendship relations, some pairs of users present repeated interactions. This is not what we want to evaluate with this metric. Instead, we want to measure the concentration of interactions on the global scale, i.e., several different actors interacting with a same one. To eliminate the effect of the repeated messages due to friendship, we rewire all interactions between the same pair of users as a set of interactions between another pair of users.

Our rewiring procedure is therefore defined as the following process: for each set of edges $s = e_1, e_2, \dots, e_x$ such as $e_1 = e_2 = \dots = e_x = (a, b)$, we replace it by another set of edges s' , such as $|s'| = |s|$ and $e'_1 = e'_2 = \dots = e'_x = (a, c)$, with $c \neq a$. Note that the out-degrees of nodes are also preserved.

3.5.3 Computation of the concentration impact

We compute the difference between the distribution of concentration in the null model and in the observed model as described in section 3.2.

3.6 Interpretation of the metrics

We have now defined three metrics, SSI, RI, and CI, having a value between 0 and 1, and that characterizes the nature of interactions between a set of users. We summarize in this section their interpretation:

- SSI represents how strong is the impact of the social structure on the interactions, how much more repeated interactions there are compared to the case in which interactions are done randomly between actors. A value of 0 means that there are not more repeated

interactions than in the random case, while a value of 1 means that none of the repeated interactions observed could be explained by random interactions. The higher the value, the more important is the impact of the social structure.

- RI represents how often we observe reciprocal interactions between actors, compared to random interactions. A value of 0 means that the observation is not distinguishable from random. The higher the value, the more actors tend to initiate interaction towards actors that also initiate interactions towards them.
- CI represents how much some actors concentrate a large fraction of the interactions. In the case of interactions made at random between users, the number of interactions received by the different actors is mostly even. In some interaction networks, on the contrary, some actors are much more popular than others, for instance, they receive an important fraction of all interactions. This metric computes how much more concentration there is in the studied dataset compared with the random case. A value of 0 means that there is no more concentration than in the random case, and a value of 1 means that none of the concentration observed could be explained by random interactions (for instance, if all interactions initiated by all actors have the same recipient).

3.7 Validation using a generative model

In order to show that our metrics do capture a property of interaction networks, independent of their size and average degree, we validate them using a generative model. We will present in detail the validation of the SSI metric.

The generative model we propose uses four parameters:

- nba : number of actors
- nbi : number of interactions
- $afpa$: average number of friends per actor
- fs : friendship strength, chosen between 0 and 1

First, we define an interaction multigraph $G = (V, E)$, with $|V| = nba$ and $E = \{\}$.

Secondly, we generate a random simple graph of friendship $GF = (V, E')$, in which nodes are the same as in G , and edges represent friendship relations. E' is generated randomly, using a power law degree generator (the distributions of in-degrees and out-degrees follow a power law of a given exponent, chosen arbitrarily as 2.5). The number of edges is chosen such as $\frac{|E'|}{|V|} = afpa$.

Finally, we generate the set of interactions E . E is composed of $nbi \times (1 - fs)$ edges chosen randomly between nodes of V , and of $nb \times fs$ edges chosen randomly among edges in E' .

fs is the parameter that corresponds to what SSI should represent. The lower the value of fs , the less the multigraph of interactions depends on the underlying friendship structure. The higher its value, the more interactions will occur according to the friendship graph. nba , nbi , $afpa$ parameters can be used to vary the structure of the multigraph of interactions, and of the friendship graph, to test their influence.

Fig. 4 presents the evolution of the computed value of SSI for fixed values of nba , nbi , and fs , while varying the value of $afpa$, which corresponds to the quantity of information we have in terms of number of interactions. We can make several observations:

- After a monotonic growth, the value of SSI stabilizes itself at a value very close to the chosen value of fs .
- The rate of change of the computed SSI slows down as $afpa$ increases, which means that a relatively limited quantity of information can provide a decent approximation of the final value.
- The size of the network and the number of friendships have an influence on the quantity of data needed to reach the stabilized value. Less friendships and a larger network make the convergence faster. This can be explained by the following phenomena: for a same number of observed interactions, more friendship relations means that we are less likely to observe unusually repeated interactions, while in a small network, the probability of observing repeated interactions between non-friend is increased.

We can therefore conclude that our approach successfully captures the impact of the social structure, at the condition of having enough data.

3.8 Analysis of results

We computed the values of social structure impact, reciprocity impact, and concentration Impact on five interaction networks issued from four different datasets. These datasets reflect different types of interactions. By comparing the results obtained on the NND dataset with the ones of other, better-understood networks, we can uncover its characteristics.

3.8.1 Description of analyzed networks

We chose datasets corresponding to different types of interactions. Below, we describe briefly each dataset, what it represents, and how we constructed the corresponding interaction network.

- The DBLP (Ley 2002) is a well-known database of scientific publications. We used a version including references between articles, as described in Tang et al. (2008). The network of interaction that we create is a network of citations. To produce it, we associate one node with each person appearing as author of a publication. Then, for each valid citation, we create a link from each of the authors of the citing article to each of the authors of the cited one, excluding self-citations.
- We use a Twitter dataset described in Toriumi et al. (2013), Remy et al. 2013), that contains between 80 and 90 % of all tweets published between Japanese users on a period of 22 days (around the Japanese earthquake and Tsunami of 2011). We kept only tweets between active users, defined as having tweeted at least

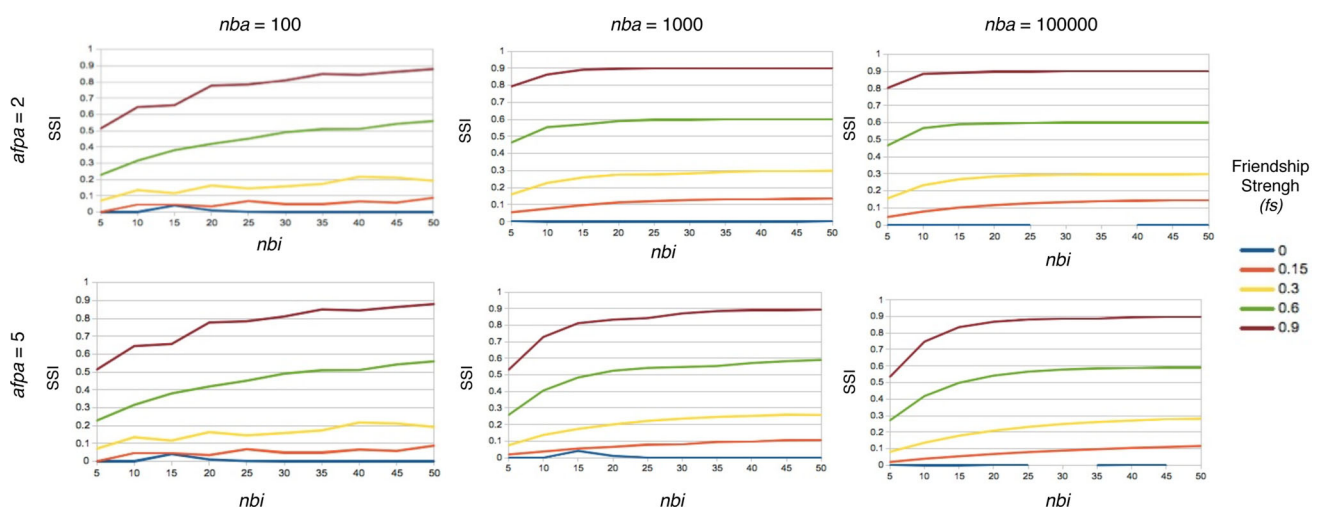


Fig. 4 Evolution of the computed value of SSI for generated graphs, varying the parameters $afpa$, nba , and fs . The computed value of SSI always reach and stabilize at the same value as the parameter fs . This confirms that the value of SSI represents the strength of friendship,

independent of the size of the network and the number of friends. However, we observe a variation in the number of observations nbi needed to reach this stabilization

a tweet and been referenced at least ten times. There are slightly above 300,000 such users. From this dataset, we extracted two networks: a network of retweets and a network of direct mentions, excluding retweets. In both networks, nodes correspond to Twitter users. In the retweet network, later called TwitterRT, we consider only tweets starting with the character chain RT@, and containing a single user name. For each of these tweets, we create an interaction link from the author of the tweet to the retweeted user. For the second network, later identified as TwitterNOTRT, we consider only the tweets starting by a mention, containing only one mention, and not containing the chain of character “RT”, characteristic of a retweet. The idea behind these two networks is that they are characteristic of two different practices: retweets are often used as a way to diffuse information, while direct mentions are often used to interact directly with a user.

- The ENRON dataset is a widely used dataset about mail communication. See Klimt and Yang (2004) for a complete description of the dataset. It contains most of the emails sent and received on the professional addresses of some of the key individuals involved in the Enron Scandal. We filtered the dataset to keep only the messages sent between the individuals who were the focus of the data collection, 154 employees of the Enron company. Nodes correspond to individuals, and each email sent from one of these persons to another is represented as a directed edge.
- The construction of the NND interaction network has already been described.

The Table 1 summarizes the number of nodes and edges of each of these communication networks, as well as the average number of interactions per actor (I/A), equivalent to the *nbi* parameter of our simulation. For all networks but NND, we provide two values for each metric. As we have seen in the validation of the SSI metric, the value of the metric can change according to the quantity of data that we consider. In real datasets, we do not have an infinite quantity of data, so we cannot always reach a point of stabilization of the value. In particular, in NND, the total

number of interactions per user is less than 14, the lowest value of all datasets. To take this limitation into account in our comparison, for all the other datasets, we propose two values: first, the value when we first reach a communication per user of 14, secondly, the value when considering all available information. We can note that, although there are modifications in the values, there is no radical change. In particular, datasets having low values compared to others keep low values, and similarly for high values. We also observe that in a few occurrences, the value decreases when considering more data. We have to stress that these being real datasets, the behaviors of users might change during the observed dataset. Both ENRON and Twitter datasets have been collected in the time of crisis, for instance, during which the behavior of users can change (Remy et al. 2013).

3.8.2 Interpretation of results

The values of social structure impact vary from network to network, but is always above 0.25, which means that at least 25 % of the concentration of interactions cannot be explained by a model of random interactions. The TwitterNOTRT network has the highest value, while the TwitterRT network has the lowest. We can propose an intuitive explanation for this observation: in the case of retweets, people are likely to retweet an information because they consider it interesting, more than because they have a particular social relationship with the author. The social relationship, such as being a follower, favor the chances to have retweets, but people are free to retweet any tweet of someone they do not know, generating random-like interactions. On the contrary, users will tend to make direct mentions of people only if they have some sort of social bond with them, as a mention is a way to directly reach the mentioned user.

For the Reciprocity Impact, we can divide the studied networks in two categories: on the one hand, NND, DBLP, and TwitterRT have low values. NND and TwitterRT in particular have values below 0.03, which means that more than 97 % of the reciprocity observed can be explained by

Table 1 Description of the five interaction network studied

Network	Nodes	Edges	I/A	SSI	RI	CI
NND	27,514	371,450	13.5	0.36	0.023	0.37
DBLP	194,079	7,940,131	40.9	0.29–0.45	0.07–0.11	0.04–0.16
TwitterRT	271,402	16,917,969	62.33	0.31–0.28	0.019–0.018	0.32–0.40
TwitterNOTRT	262,545	17,719,946	67.5	0.75–0.88	0.63–0.66	0.0037–0.021
ENRON	155	9646	62.2	0.55–0.51	0.30–0.31	0.06–0.001

Nodes: number of nodes in the network. Edges: number of edges in the network

I/A interactions per actor (degree), SSI social structure impact, RI reciprocity impact, CI concentration impact

a random model. On the other hand, TwitterNOTRT and the ENRON datasets have much higher values. We can therefore make a distinction between the networks with high reciprocity and the others.

For the Concentration Impact, we can make a similar distinction between, on the one hand, NND and TwitterRT, which have a high value (around 30 %) and the three others that have smaller ones. Observing a high value of concentration for the retweets is not surprising, as it is known that a minority of users attracts a large fraction of all retweets. Low scores for the TwitterNOTRT and ENRON datasets are not surprising either, as they correspond to interactions that are more interpersonal, with less possibilities of aggregation.

From these observations, we can propose to classify these interaction networks into three categories:

- ENRON and TwitterNOTRT have a high SSI, high RI, and low CI. These profiles correspond to interpersonal communications, on which users tend to communicate much with other users they know, in an interactive manner.
- TwitterRT and NND have a lower—but still high—SSI, a low RI, and a high CI. These profiles correspond to diffusion networks: most of the communications do not take place between “equals”, but, rather, a small proportion of users attract most of the interactions, and do not reciprocate these interactions.
- DBLP has a different profile, with an SSI similar to the diffusion networks, a low RI and a low CI. The interactions are far from being random, but this is due neither to the major role of a few key users, nor to the strong influence of interpersonal communications. Although we do not investigate this case more in depth in this article, we can propose as an explanation that the bias in interactions observed is rather a “field” bias, i.e., most of the repeated interactions between users can be explained by the fact that authors work on the same topics, on the same scientific questions, and are therefore more likely to cite each other, even though this corresponds neither to interpersonal communications nor to the attraction of a few.

3.8.3 Classification of the interactions in NND

Using these indicators, we have been able to characterize the type of interactions happening on NND. We have evaluated that the impact of the social structure was as important as in other well-studied networks such as the DBLP scientific citation network or the Twitter retweet network. We have also shown that the nature of the social structure was comparable to the one happening in the retweeter network of Twitter, i.e., a diffusion of

information network, rather than interpersonal interactions such as the ones happening with mentions in Twitter, or in the ENRON email dataset. It is also different from the DBLP citation network, despite a likeliness in nature (publication of creations based on previously published creations), because of a much higher importance of the interactions with a minority of key users.

3.9 Applications

Through the application of this approach for characterizing the structure of interactions to several datasets, we have seen that it is a useful tool to grasp the differences between datasets. By the mean of the three proposed metrics, we can understand how do users interact in a dataset we do not know, and find similarities with other ones. This method can therefore be applied to gain insights into new datasets, and to better understand the nature of human communication.

4 Roles in the cooperation process

In the second part of this article, we focus on a different aspect of interactions. Whereas the first approach was focused on interactions between actors, without considering the published contents, this section studies the relations between these published contents, if they exist. To do so, we define possible roles that can be played by these published contents.

The identification of roles is a common question in collective behaviors involving interactions between individuals. In sociology, the role theory aims at assigning social roles to actors, according to their positions and actions relatively to the human society.

In network analysis, several role categorization have been proposed, such as *Bridge*, *Gateway*, *Hub*, and *Loner* in Chou and Suzuki (2010), or *Ambassadors*, *Bridges*, *Loners*, and *Big fish* in Scripps et al. (2007).

Roles have also been proposed in the case of diffusion processes, in particular the decomposition in *Idea starters*, *Amplifiers*, and *Transmitters* (Cazabet et al. 2014; Tinati et al. 2012).

These roles are useful to qualify users, static elements through which some dynamic information transit, and are well suited to study diffusion of information processes such as Twitter retweets, or communications among users in general. But some online social networks are not centered on communication, but rather on cooperation. Youtube, Wikipedia, GitHub, or NND are, for instance, more centered on production and sharing than on communication and diffusion, although users play, of course, an important role too. The process of scientific research, with its authors and published articles, is of the same nature. For these categories of

networks, it is useful to attribute a role to the pieces of data themselves. A role can consequently be attributed to authors based on the type of creation they publish.

We define roles based on the topological structure of the network $G = (V, E)$ of references between productions. In the case of NND, V corresponds to the set of videos in our dataset. The set of edges E is defined such as $(a, b) \in E \iff$ the video a references the video b (the source of the reference is also the source of the edge). Note that the orientation of edges is somewhat reversed compared to a diffusion network: the source of an edge is taking some information from its target.

Some of these definitions use a common threshold value, $t_{\text{Influential}}$, which is chosen according to the dataset and the level of granularity that we want to get, and which corresponds to a minimal number of videos needed to be considered of a given role. In the rest of our analysis, we use $t_{\text{Influential}}$, which means that a video will need to be referenced at least ten times to be considered influential, for instance.

4.1 Definition of roles

This section defines the roles of videos. They are organized in direct roles, relative roles, and indirect roles.

4.1.1 Direct roles

We start by defining three mutually exclusive roles, based on well-established network properties. These roles can be defined using only the direct neighbors of a node.

Original creations are defined as nodes making no references to any other node, but being referenced at least once, i.e., sink in graph theory.

$$n \in \text{OriginalC} \iff d_{\text{in}}(c) > 0 \wedge d_{\text{out}}(c) = 0$$

Dead ends are defined as nodes that reference at least one other video but are never referenced, i.e., source in graph theory.

$$n \in \text{DeadEnd} \iff d_{\text{in}}(c) = 0 \wedge d_{\text{out}}(c) > 0$$

Influential creations are nodes that are cited more than a threshold $t_{\text{Influential}}$. This threshold can be equal to a chosen value, a fraction of the most referenced nodes, or a function such as the average number of references.

$$n \in \text{IC} \iff d_{\text{in}}(c) > = t_{\text{Influential}}$$

4.1.2 Relative roles

These roles are not intrinsic to the nodes but defined relatively to another one. They are a preliminary step to the computation of indirect roles.

Simple variants: n is a simple variant of a creation n' if it references n' , and uses nothing new compared to n' .

$$n \in \text{SV}(n') \iff n_{\text{out}}(n) \subset \{\text{succ}(n') \cup \{n'\}\},$$

where $\text{succ}(v)$ is the set of all successors (direct and indirect) of v .

Complex variants: n is a complex variant of a creation n' if it references both n' and at least one other simple variant or complex variant of n' .

$$n \in \text{CV}(n') \iff n' \in n_{\text{out}}(n) \wedge (\exists p \in n_{\text{out}}(n) | p \in \text{CV}(n') \vee p \in \text{SV}(n')),$$

Exploiting creations: n is an exploiting creation of n' if it is not a complex variant of n' , but references at least one creation that is not a successor of n' (contrary to simple variants). Because these definitions are exclusive and complementary, we can define exploiting creations as:

$$n \in \text{EC}(n') \iff n' \in n_{\text{out}}(n) \wedge n \notin \text{CV}(n') \wedge n \notin \text{SV}(n')$$

4.1.3 Indirect roles

These roles are defined by considering the direct roles of nodes and the relative roles of their neighbors.

Local inspirers are influential creations that inspire a large number of simple variants.

$$n \in \text{LI} \iff n \in \text{IC} \wedge |\{(v, n) \in E | v \in \text{SV}(n)\}| > = t_{\text{Influential}}$$

with $t_{\text{Influential}}$ is the same threshold as chosen to define an influential creation.

Global inspirers are influential creations that inspire a large number of complex variants.

$$n \in \text{GI} \iff n \in \text{IC} \wedge |\{(v, n) \in E | v \in \text{CV}(n)\}| > = t_{\text{Influential}}$$

with $t_{\text{Influential}}$ is the same threshold as chosen to define an influential creation.

Building blocks are influential creations that are used by a large number of exploiting creations.

$$n \in \text{BB} \iff n \in \text{IC} \wedge |\{(v, n) \in E | v \in \text{EC}(n)\}| > = t_{\text{Influential}}$$

with $t_{\text{Influential}}$ is the same threshold as chosen to define an Influential Creation.

Aggregators are videos n which are making references to at least two videos with no common successors in the graph. The aim of this definition is to select videos that use independent sources of inspiration.

$$n \in \text{Agg} \iff \exists a, b | (n, a) \in V \wedge (n, b) \in V \wedge \text{succ}(a) \cap \text{succ}(b) = \emptyset$$

where $\text{succ}(v)$ is the set of all successors (direct and indirect) of v .

An example of a small network on which these roles have been attributed is represented in Fig. 5.

In Fig. 6, we present subnetworks issued from our dataset, using the same legend for colors. These networks are obtained by the following process:

1. Selection of an initial node
2. Collection of the subnetwork composed of all nodes referencing the initial node and all edges among them
3. Removal of all transitive edges in the obtained subnetwork, in order to keep the graph simple enough to be visualized.

4.2 Demography of roles

We can study, for the roles we have defined, how common they are, globally and by categories of creation.

4.2.1 Demography of direct roles

In Table 2, we present the distribution of direct roles. We can see that nearly three-quarters of creations are Dead Ends—not surprising, given the power law distribution of degrees—while 16 % are Original Creations. But we can also note large disparities between categories.

Singing videos are the most likely to be Dead Ends, while it is much more unlikely than average for **OriginalMusic** and **Animation** videos, in particular.

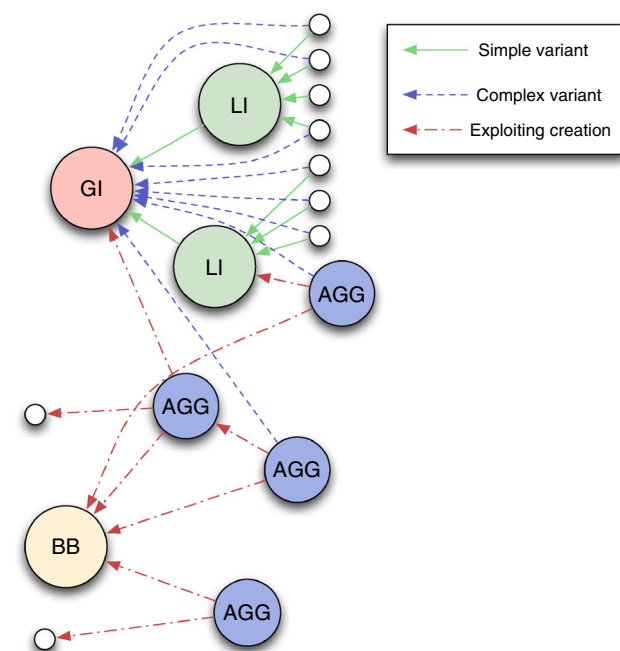


Fig. 5 Schema of indirect roles on a toy example, with the threshold $t_{\text{Influential}} = 4$

We can see that **OriginalMusic** are the most likely to be Original Creations, while it is rarer than average for **Singing** videos in particular.

Finally, we can observe that influential creations are rare for all categories, but large variations can nevertheless be observed, **OriginalMusic** and **Animation** being the most likely to be influential, while **Singing**, **Mashups**, and **MAD** videos are the less likely to be.

4.2.2 Demography of indirect roles

Although influential creations represent only 1.3 % of videos, they play very important roles in the cooperation process. We know by definition that Local Inspirers, Global Inspirers, and Building Blocks can be found among these influential creations. There is a total of 9007 influential videos, with $t_{\text{Influential}} = 10$, and Table 3 represents the distribution of each role for each category of creation having more than 50 influential creations. We can observe that some categories have clearly pronounced profiles.

All categories taken together, more than half of influential creations are local Inspirers, while global inspirers and building blocks have comparable prevalence.

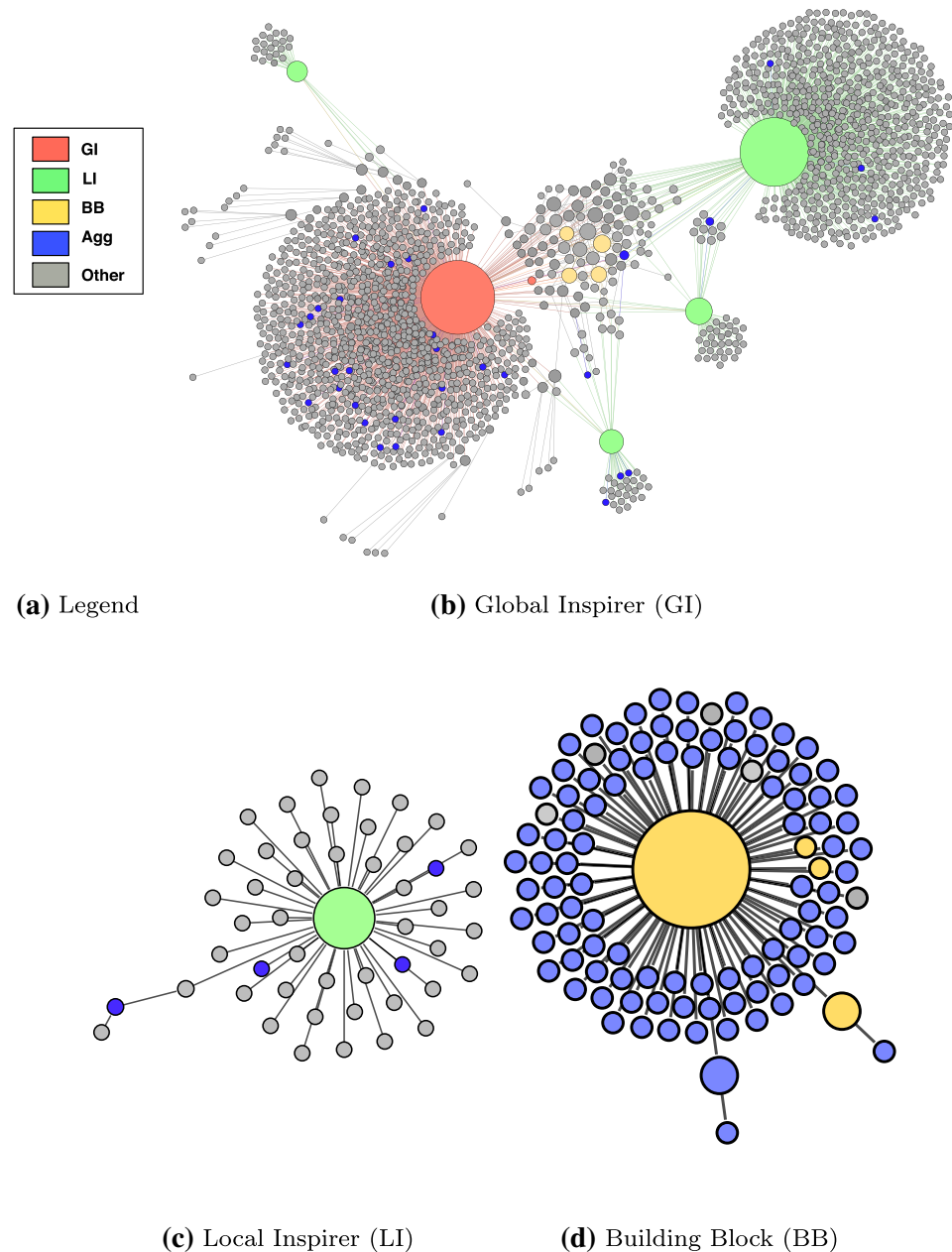
- **OriginalMusic** has high LI, high GI, and low BB
- **CG3D** has low LI and GI, but high BB
- **Voice**, **MusicalPerformance**, and **Animation** have high LI, low GI, and low BB
- **Dance** has average LI and high BB

These findings are coherent with what we observe in the dataset. **CG3D**, **Dance**, and **Picture** videos have high BB, and they often contain elements that can be used in unrelated videos. For instance, **CG3D** videos are sometimes demonstrations of how to use a 3D tool, or are demonstrations of 3D models that others are free to use in their creations. Some **Dance** videos also demonstrate choreographies that can be used for unrelated music. Finally, elements of **Picture** videos are also easy to reuse in different contexts.

On the contrary, **OriginalMusic** are barely used to create unrelated videos, but are the most inspiring on the global level. Indeed, most of the cooperation processes are based on an initial **OriginalMusic**. These videos are also important LI, because many videos are simple variations of them, such as someone singing it in a different manner.

Voice, **MusicalPerformance**, and **Animation** are powerful LI inspirers, but generate none or very few Global inspiration. This is also easily observed in the dataset. Influential **Voice** videos for instance are in majority “karaoke” versions of a famous song, i.e., with subtitles and no

Fig. 6 Subnetworks of the NND citation network centered on an Influential Creation of each role. Different patterns are observed. The sizes of the nodes are proportional to their degrees



lyrics. As a consequence, they are massively reused by singers, but do not inspire more complex creations.

Finally, it is interesting to note that **CG3D** is the only category with barely any LI. It is because most other famous videos attract a lot of simple “copies”, while it is not the case for **CG3D** videos.

Aggregator roles are studied in Table 4. Not surprisingly, **Mashups** are the most likely to be aggregators, with more than 60 % of them recognized as such. **CG3D**, and to a lesser extent **Picture** and **Dance** are also frequent aggregators. **OriginalMusic** videos, being often original creations, are very unlikely to be aggregators.

4.2.3 Generalization to other types of cooperation

The roles we propose are meaningful in NND, but are also generic enough to make sense in different cooperative creation processes. If we take as an example the domain of scientific publications, modeled by the network of citations among papers, we can assimilate Building Blocks to articles that propose a tool, a method, or that present a dataset, and which are therefore cited for this particular element. Aggregators would correspond to articles of review. Global inspirers would be seminal articles that inspire a new field of research. Finally, local inspirers are publications that are

Table 2 Ratio of each direct role among videos of each category, and overall

	Dead ends	Original creations	Influential creation
Overall	0.74	0.16	0.013
Animation	0.32	0.36	0.060
CG3D	0.66	0.15	0.023
Dance	0.80	0.07	0.014
MAD	0.44	0.46	0.000
Mashups	0.75	0.11	0.004
Movie	0.39	0.26	0.010
Music	0.51	0.28	0.000
MusicalPerformance	0.67	0.22	0.010
OriginalMusic	0.02	0.93	0.053
Picture	0.76	0.10	0.011
Singing	0.91	0.02	0.003
VocaloidVoice	0.53	0.28	0.015
Voice	0.71	0.08	0.029

Most remarkable features appear in bold

Table 3 Percentage of all Influential videos which are of a given role, by category and overall

Category	LI	GI	BB
Overall	0.59	0.16	0.14
CG3D	0.08	0.01	0.64
Dance	0.48	0.04	0.25
Voice	0.74	0.00	0.05
OriginalMusic	0.73	0.24	0.02
MusicalPerformance	0.71	0.00	0.06
Singing	0.45	0.12	0.08
Animation	0.68	0.04	0.04

Values above average are in bold

imitated or inspire similar articles, but without having a large-scale influence on their field. We can note however that the network definitions used in the context of NND would not give relevant results in the case of scientific citations. The main reason is probably that the meaning of citations in science is very different to the meaning of references in NND. In NND, only the actively used sources are referenced, while in science, we tend to cite many articles that are only loosely related to the very contribution of the article. Different methods, having a probabilistic approach or using machine learning, would be necessary to uncover these roles in such a network, if feasible by considering only the topology of the citation network.

4.3 Creators and roles of videos

The method we propose attributes roles to productions of the social network. As many researches and applications focus on users, we subsequently explore the relation between users and the roles of the videos they produce. For

Table 4 Ratio of videos being aggregators, by category and overall

Category	AGG
Overall	0.10
Animation	0.09
CG3D	0.40
Dance	0.16
MAD	0.07
Mashups	0.61
Movie	0.09
Music	0.08
MusicalPerformance	0.06
OriginalMusic	0.01
Picture	0.17
Singing	0.04
VocaloidVoice	0.10
Voice	0.05

Highest and lowest values are in bold

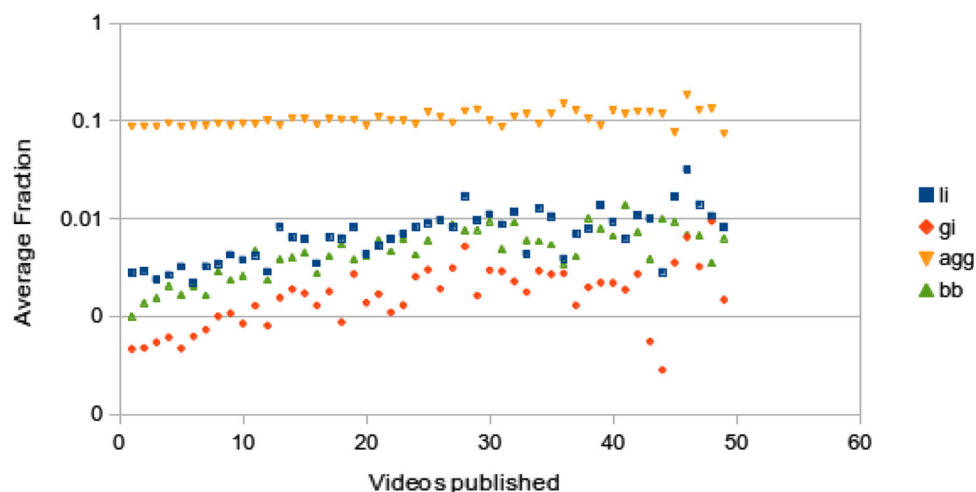
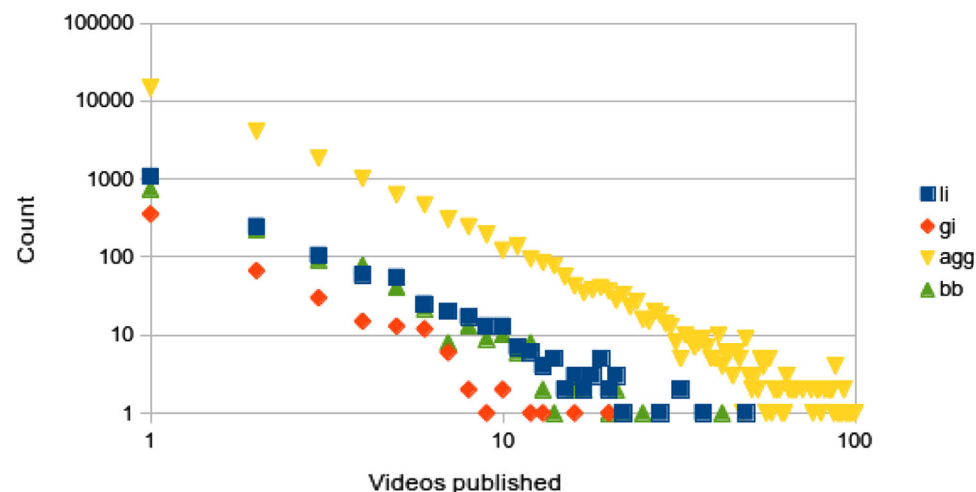
each user, we generate seven indicators: the total number of references **MADe** by the videos he published, the total number of references his videos received, the number of videos he published, and the number of videos he published for each indirect role (LI,GI,AGG,BB).

Table 5 presents the Pearson correlation coefficients between these indicators. We can observe that some present a very low correlation, such as the number of videos published and the number of citations received. Some characteristics are more correlated, in particular the number of citations received with the number of GI videos published (0.8) and the number of references **MADe** with the number of AGG videos published (0.77). These two correlations are not a surprise, and confirm that the

Table 5 Pearson correlation coefficients between several characteristics of users, namely the number of references in the videos they published, the number of references the video they published

	Ref made	Ref received	Videos	LI	GI	AGG	BB
Ref made	1	0.02	0.40	0.04	0.01	0.77	0.05
Ref received		1	0.16	0.53	0.80	0.04	0.32
#videos			1	0.28	0.13	0.47	0.20
LI				1	0.49	0.03	0.27
GI					1	0.02	0.30
AGG						1	0.13
BB							1

Remarkably high and low values are highlighted

Fig. 7 For each role, detail of the correlation between the number of videos published by author and the proportion of these videos having this role. Users with the highest number of videos published also have the highest proportions of videos of each role**Fig. 8** Distributions of the number of videos, for each role. They present a long tail, which means that a few users publish exceptionally large quantities of videos of each role

identified roles do capture efficiently the properties of the network.

Fig. 7 explores in more detail the relation between the number of videos published by author and the average

received, the number of video they published, the number of local inspirers, global inspirers, aggregators, and building blocks among the videos they published

proportion of their videos having a given role. We observe that the average proportion of aggregator videos is mostly independent of the number of videos published, staying roughly around 10 %. On the contrary, the probability of

publishing LI, GI, and BB videos increases with the number of videos published. The most active users are therefore also the ones whose videos have the most important roles in the creation process.

Fig. 8 presents, for each role, the distribution of the number of videos of this role published by user. The distributions have a power law profile, with a large fraction of users publishing only one video of a given role, while a few users publish exceptionally large numbers of it. There is a concentration of the production of videos having well-defined roles.

5 Conclusion

The contribution of this article is twofold. On the one hand, we uncovered important characteristics of a unique dataset of massive scale artistic cooperation in an online social network. On the other hand, we proposed two generic methods, using network properties, which can be applied to different datasets, to better understand and compare them.

The first contribution is the uncovering of the nature of the cooperation in the Nico Nico Douga Online Social Network. This network is unique because it represents a rare case of artistic cooperation conducted in a large scale, with a large quantity of data being available. This mechanism of creation based on previous creations is ubiquitous on the internet, for instance in the emergence of *memes*. However, this creation process has rarely been studied, due to the difficulty of collecting data on the source of inspirations. By studying this process in NND, we discovered that this mechanism was comparable on many aspects to information diffusion: weak reciprocity, high importance of a few key inspirers. Using roles, we have also shown that different sort of contributions with different profiles exist, that we could correlate in our dataset with the categories of videos. Furthermore, this method allows us to identify key contributions to the cooperation process, more precisely than using only the in- and out-degrees of nodes. Finally, we have shown that it was also possible to identify the roles of users by considering the roles of the videos they publish.

The second contribution is the proposition of generic methods for studying interactions. Both of the methods we introduced can be reused easily on different datasets, because they do not consider the content or the nature of the interactions, but only the network that they form. The characterization of the nature of the interactions in particular, through the SSI, CI, and RI metrics, can be applied to any sort of interaction dataset, as we have shown by studying sources as varied as email exchange, scientific citations, and Twitter messaging. The results show a clear differentiation between them, compatible with our

knowledge of their nature. We provide an open source code² to compute these values on any such network.

The second method, the attribution of roles to published contents, is more specific to cooperative processes, and might need adaptations to work on different networks, but we believe nevertheless that the roles defined are generic enough to make sense in other contexts.

Acknowledgments This work was supported in part by OngaCREST, CREST, JST

References

- Amaral LAN, Scala A, Barthelemy M, Stanley HE (2000) Classes of small-world networks. *Proc Natl Acad Sci* 97(21):11149–11152
- Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on twitter. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, pp 65–74
- Cazabet R, Takeda H, Hamasaki M, Amblard F (2012) Using dynamic community detection to identify trends in user-generated content. *Soc Netw Anal Min* 2:361–371
- Cazabet R, Takeda H (2014) Understanding mass cooperation through visualization. In: *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, pp 206–211
- Cazabet R, Pervin N, Toriumi F, Takeda H (2014) Using network properties to analyze users' role in twitter in time of crisis. In: *28th Annual Conference of the Japanese Society for Artificial Intelligence*, Matsuyama, 12–15 May 2014
- Cha M, Haddadi H, Benevenuto F, Gummadi PK (2010) Measuring user influence in twitter: the million follower fallacy. *ICWSM* 10:10–17
- Chou BH, Suzuki E (2010) Discovering community-oriented roles of nodes in a social network. In: *Data warehousing and knowledge discovery*. Springer, pp 52–64
- Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. *SIAM Rev* 51(4):661–703
- Hamasaki M, Takeda H, Hope T, Nishimura T (2009) Network analysis of an emergent massively collaborative creation community: how can people create videos collaboratively without collaboration? In: *Third International AAAI Conference on Weblogs and Social Media*
- Hamasaki M, Takeda H, Nishimura T (2008) Network analysis of massively collaborative creation of multimedia contents: case study of hatsune miku videos on nico nico douga. In: *Proceedings of the 1st international conference on Designing interactive user experiences for TV and video*. ACM, pp 165–168
- Hamasaki M, Goto M (2013) Songrium: a music browsing assistance service based on visualization of massive open collaboration within music content creation community. In: *Proceedings of the 9th International Symposium on Open Collaboration*. ACM, p 4 (2013)
- Klimt B, Yang Y (2004) The enron corpus: a new dataset for email classification research. In: *Machine learning: ECML 2004*. Springer, pp 217–226
- Leskovec J, Lang K, Dasgupta A, Mahoney M (2009) Community structure in large networks: natural cluster sizes and the absence of large well-defined clusters. *Internet Math* 6(1):29–123

² On the webpage of the first author

- Ley M (2002) The dblp computer science bibliography: evolution, research issues, perspectives. In: *String Processing and Information Retrieval*. Springer, pp 1–10
- Nakamura S, Shimizu M, Tanaka K (2008) Can social annotation support users in evaluating the trustworthiness of video clips? In: *Proceedings of the 2nd ACM workshop on Information credibility on the web*. ACM, pp 59–62
- Remy C, Pervin N, Toriumi F, Takeda H (2013) Information diffusion on twitter: everyone has its chance, but all chances are not equal. In: *Signal-Image Technology and Internet-Based Systems (SITIS), 2013 International Conference on*, IEEE. pp 483–490
- Scripps J, Tan PN, Esfahanian AH (2007) Node roles and community structure in networks. In: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. ACM, pp 26–35
- Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z (2008) Arnetminer: extraction and mining of academic social networks. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp 990–998
- Tinati R, Carr L, Hall W, Bentwood J (2012) Identifying communicator roles in twitter. In: *Proceedings of the 21st international conference companion on World Wide Web*. ACM, pp 1161–1168
- Toriumi F, Sakaki T, Shinoda K, Kazama K, Kurihara S, Noda I (2013) Information sharing on twitter during the 2011 catastrophic earthquake. In: *Proceedings of the 22nd international conference on World Wide Web companion, International World Wide Web Conferences Steering Committee*, pp 1025–1028
- Yang J, Counts S (2010) Predicting the speed, scale, and range of information diffusion in twitter. *ICWSM* 10:355–358